Comparison of single distribution and mixture distribution models for modelling LGD

Jie Zhang and Lyn C Thomas
Quantitative Financial Risk Management Centre,
School of Management, University of Southampton

Abstract
Estimating Recovery Rate and Recovery Amount has taken a more importance in consumer credit because of both the new Basel Accord regulation and the increase in number of defaulters due to the recession.
We examine whether it is better to estimate Recovery Rate (RR) or Recovery amounts. We use linear regression and survival analysis models to model Recovery rate and Recovery amount, thus to predict Loss Given Default (LGD) for unsecured personal loans. We also look at the advantages and disadvantages of using single distribution model or mixture distribution models for default.

Key words: Recovery Rate, Linear regression, Survival analysis, Mixture distribution

1. Introduction

The New Basel Accord allows a bank to calculate credit risk capital requirements according to either of two approaches: a standardized approach which uses agency ratings for risk-weighting assets and internal ratings based (IRB) approach which allows a bank to use internal estimates of components of credit risk to calculate credit risk capital. Institutions using IRB need to develop methods to estimate the following components for each segment of their loan portfolio:
– PD (probability of default in the next 12 months);
– LGD (loss given default);
– EAD (expected exposure at default).

Modelling PD, the probability of default has been the objective of credit scoring systems for fifty years but modelling LGD is not something that had really been addressed in consumer credit until the advent of the Basel regulations. Modelling LGD is more difficult than modelling PD. There are two main reasons: first, data may be censored (debts still being paid) because of long time scale of recovery. Linear regression does not deal that well with censored data and even the Buckley-James approach does not cope well with this form of censoring. Second, debtors' different view about default leads to different repayment patterns. For example, some people deliberately do not want to repay; some people can not repay, but there will be different reasons for this inability to repay and one model can not deal with them. Survival analysis though can handle censored data, and segmenting the whole default population is helpful to modelling LGD for defaulters with different reasons for defaulting.

Most LGD modelling is in the corporate lending market where LGD (or its opposite Recovery Rate RR, where RR=1-LGD), was needed as part of the bond pricing formulae. Even there, until fifteen years ago LGD was assumed to be a deterministic value obtained from a historical analysis of bond losses or from bank work out

experience (Altman et al 1977). Only when it was recognised that LGD was needed for the pricing formula and that one could use the price of non defaulted risky bonds to estimate the market's view of LGD were models of LGD developed. If defaults are rare in a particular bond class then it is likely the LGD got from the bond price is essentially a subjective judgment by the market. The market also trades defaulted bonds and so one can get directly the market values of defaulted bonds (Altman and Eberhart 1994). These market values or implied market values of Loss Given Default were used to build regression models that related LGD to relevant factors, such as the seniority of the debt, country of issue, size of issue and size of firm, industrial sector of firm but most of all to economic conditions which determined where the economy was in relation to the business cycle. The most widely used model is the Moody's KMV model, LossCalc (Gupton 2005), it transforms the target variable into normal distribution by a Beta transformation; then regresses the transformed target variable on a few characteristics, and then transforms back the predicted values to get the LGD prediction. Another popular model, Recovery Ratings, was created by Standard & Poor's Ratings Services (Chew and Kerr 2005); it classifies the loans into 6 classes which cover different recovery ranges. Descriptions of the models are given in several books and reviews (Altman, Resti, Sironi 2005, De Servigny and Oliver 2004, Engelmann and Rauhmeier 2006, Schuermann 2004).

Such modelling is not appropriate for consumer credit LGD models since there is no continuous pricing of the debt as is the case on the bond market. The Basel Accord (BCBS 2004 paragraph 465) suggests using implied historic LGD as one approach in determining LGD for retail portfolios. This involves identifying the realised losses (RL) per unit amount loaned in a segment of the portfolio and then if one can estimate the default probability PD for that segment, one can calculate LGD since RL=LGD.PD. One difficulty with this approach is that it is accounting losses that are often recorded and not the actual economic losses, which should include the collection costs and any repayments after a write-off. Also since LGD must be estimated at the segment level of the portfolio, if not at the individual loan level there is often insufficient data in some segments to make robust estimates.

The alternative method suggested in the Basel Accord is to model the collections or work out process. Such data was used by Dermine and Neto de Carvalho (Dermine and Neto de Carvalho 2006) for bank loans to small and medium sized firms in Portugal, but they used a regression approach, albeit a log-log form of the regression to estimate the data.

The idea of using the collection process to model LGD was suggested for mortgages by Lucas (2006). The collection process was split into whether the property was repossessed and the loss if there was repossession. So a scorecard was built to estimate the probability of repossession where Loan to Value was key and then a model used to estimate the percentage of the estimated sale value of the house that is actually realised at sale time. For mortgage loans, a one-stage model, was build by Qi and Yang (2007). They modelled LGD directly, and found LTV (Loan to Value) was the key variable in the model and achieved adjusted R square of 0.610, but only a value of 0.15 without it

For unsecured consumer credit, the only approach is to model the collections process, as there is no pricing mechanism for the debt, equivalent to the bond price for

corporate debt. Moreover, there is no security to be repossessed. The difficulty is that the Loss Given Default, or the equivalent Recovery Rate, depends both on the ability and the willingness of the borrower to repay, and on decisions by the lender on how vigorously to pursue the debt. This is identified at a macro level by Matuszyk et al (2007), who use a decision tree to model whether the lender will collect in house, use an agent on a percentage commission or sell off the debts, - each action putting different limits on the possible LGD. If one concentrates only on one mode of recovery in house collection for example, it is still very difficult to get good estimates. Matuszyk et al (2007) look at various versions of regression, while Bellotti and Crook (2009) add economic variables to the regression. Somers and Whittaker (2007) suggest using quantile regression, but in all cases the results in terms of R-square are poor - between 0.05 and 0.2. Querci (2005) investigates geographic location, loan type, workout process length and borrower characteristics for data from an Italian bank, but concludes none of them is able to explain LGD though borrower characteristics are the most effective.

In this paper, we use linear regression and survival analysis models to build predictive models for recovery rate, and hence LGD. Both single distribution and mixture distribution models are built to allow a comparison between them. This analysis will give an indication of how important it is to use models which cope well with censored debts and also whether mixed distribution models give better predictions than single distribution model.

The comparison will be made based on a case study involving data from an in house collections process for personal loans. This consisted of collections data on 27K personal loans over the period from 1989 to 2004. In section two we briefly describe the theory of linear regression and survival analysis models. In section three we explain the idea of mixture distribution models. In section four we build single distribution models using linear regression and survival analysis, while in section five we create mixture distribution models, thus the comparison can be made and the results are discussed. In section 6 we summarise the conclusion obtained.

2 Single distribution models

2.1 Linear regression model

Linear regression is the most obvious predictive model to use for recovery rate (RR) modelling, and it is also widely used in other financial area for prediction. Formally, linear regression model fits a response variable y to a function of regressor variables $x_1, x_2, ..., x_m$ and parameters. The general linear regression model has the form

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_m x_m + \varepsilon \qquad (2.1)$$

Where in this case

- y is the recovery rate or recovery amount
- $\beta_0, \beta_1, ... \beta_m$ are unknown parameters
- $x_1, x_{2,}, ..., x_m$ are independent variables which describe characteristics of the loan or the borrower
- ε is a random error term.

In linear regression, one assumes that the mean of each error component (random variable ε) is zero and each error component follows an approximate normal distribution. However, the distribution of recovery rate tends to be bathtub shape, so the error component of linear regression model for predicting recovery rate does not satisfy these assumptions.

2.2 Survival analysis models

Survival analysis concepts

Normally in survival analysis, one is dealing with the time in that an event occurs and in some cases the events have not occurred and so the data is censored. In our recovery rate approach, the target variable is how much has been recovered before the collection's process stops, where again in some cases, collection is still under way, so the debt is censored.

The debts which were written off are uncensored events; the debts which have been paid off or still being paid are censored events, because we don't know how much more money will be paid or could have been paid. If the whole loan is paid off, we have to treat this to be a censored observation, as in some cases, the recovery rate greater than 1. If one assumes recovery rate must never exceed 1, then such observations are not censored.

Suppose T is the random variable (defined as RR in this case) with probability density function f. If an observed outcome, t of T, always lies in the interval $[0, +\infty)$, then T is a survival random variable. The cumulative density function F for this random variable is

$$F(t) = P(T \le t) = \int_0^t f(u)du \tag{2.2}$$

The survival function is defined as:

$$S(t) = P(T > t) = 1 - F(t) = \int_t^\infty f(u)du \tag{2.3}$$

Likewise, given S one can calculate the probability density function, f(u),

$$f(u) = -\frac{d}{du}S(u) \tag{2.4}$$

The hazard function is an important concept in survival analysis because it models imminent risk. The hazard function is defined as the instantaneous rate of failure at any time, t, given that the individual has survived up to that time,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t \,|\, T \ge t)}{\Delta t} \tag{2.5}$$

The hazard function can be expressed in terms of the survival function,

$$h(t) = \frac{f(t)}{S(t)}, \quad t > 0 \tag{2.6}$$

Rearranging, we can also express the survival function in terms of the hazard,

$$S(t) = e^{-\int_0^t h(u)du} \tag{2.7}$$

Finally, the cumulative hazard function, which relates to the hazard function, $h(t)$,

$$H(t) = \int_0^t h(u)du = \ln S(t) \tag{2.8}$$

is widely used.

It should be noted that f, F, S, h and H are related, and only one of the function is needed to be able to calculate the other four.

There are two types of survival analysis models which connect the characteristics of the loan to the amount recovered – accelerated failure time models and Cox proportional hazards regression.

Accelerated failure time models
In an accelerated failure time model, the explanatory variables act multiplicatively on the survival function, they either speed up or slow down the rate of failure. If g is a positive function of x and S_0 is the baseline survival function then an accelerated failure model can be expressed as

$$S_x(t) = S_0(t \cdot g(x)) \tag{2.9}$$

Where the failure rate is speed up where $g(x) < 1$. by differentiating (2.9), the associated hazard function is

$$h_x(t) = h_0[tg(x)]g(x) \tag{2.10}$$

For survival data, accelerated failure models are generally expressed as a log-linear model, which occurs when $g(x) = e^{\beta^T x}$. Note here that if $\beta^T x = 0$ then g=1. After taking the logarithm of both sides,

$$\log_e T_x = \mu_0 + \beta^T x + \sigma Z \tag{2.11}$$

where Z is a random variable with zero mean and unit variance. The parameters, β, are then estimated through maximum likelihood methods. As a parametric model, Z is often specified as the Extreme Value distribution, which corresponds to Y having an Exponential, Weibull, Log-logistic or other types of distribution. When building accelerated failure models, the type of distribution for dependent variable has to be specified.

Cox proportional hazards regression
Cox (1972) proposed the following model

$$h(t;x) = e^{(\beta^T x)} h_0(t) \tag{2.12}$$

Where β is a vector of unknown parameters, x is a vector of covariates and $h_0(t)$ is called the baseline hazard function.

The advantage of this model is that we do not need to know the parametric form of $h_0(t)$ to estimate β, and also the distribution type of dependent variable does not need to be specified. Cox (1972) showed that one can estimate β by using only rank of failure times to maximise the likelihood function.

3 Mixture distribution models
Models may be improved by segmenting population and building different models for each segment, because some subgroups maybe have different features and distributions. For example, small and large loans have different recovery rates, long established customers have higher recovery rate than relatively new customers (the latter may have high fraudulent elements which lead to low RR), and recovery rate of house owners is higher than that of tenets (because the former has more assets which may be realisable). And also, different segments maybe have different distributions

for dependent variable, and accelerated failure time model can fit different distributions into the model, thus modelling results maybe improved.

The development of finite mixture (FM) models dates back to the nineteenth century. In recent decades, as result of advances in computing, FM models proved to offer powerful tools for the analysis of a wide range of research questions, especially in social science and management (Dias, 2004). A natural interpretation of FM models is that observations collected from a sample of subjects arise from two or more unobserved/unknown subpopulations. The purpose is to unmix the sample and to identify the underlying subpopulations or groups. Therefore, the FM model can be seen as a model-based clustering or segmentation technique (McLachlan and Basford, 1998; Wedel and Kamakura, 2000).

In order to investigate different features and distributions in subgroups, we model the recovery rate by segmenting first. A classification tree model could be built at first to generate a few segments with different features. Recovery rate is the target variable in the classification tree and we try to separate the whole population into a few segments which have different average recovery rate. Then, linear regression and survival models could be built for each segment, thus mixture distribution models can be created.

4 Case Study – Single distribution model

4.1 Data

The data in the project is a default personal loan data set from a UK bank. The debts occurred between 1987 and 1999, and the repayment pattern was recorded until the end of 2003. In total 27278 debts were recorded in the data set, of which, 20.1% debts were paid off before the end of 2003, 14% debts were still being paid, and 65.9% debts were written off beforehand. The range of the debt amount was from £500 to £16,000, 78% debts are less than or equal to £5,000 and only 3.6% of them are greater than £8,000. Loans for multiples of thousands of pound are most frequent, especially 1000, 2000, 3000 and 5000. Twenty one characteristics about the loan and the borrower were available in the data set such as the ratio of the loan to income, employment status, age, time with bank, and purpose and term of loan.

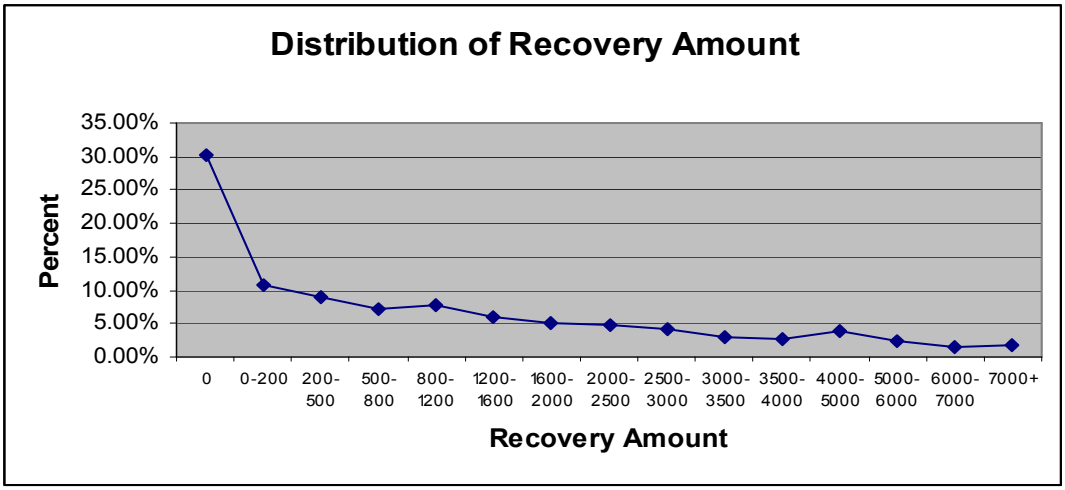

Figure (1): Distribution of Recovery Amount in the data set

The recovery amount is calculated as:

default amount – last outstanding balance (for non-write off loans)

OR default amount – write off amount (for write off loans)

The distribution of recovery amount is given in Figure 1, ignoring debts that are still being repaid but this graph could be misleading as it cannot describe the original debt.

The recovery rate $\frac{\text{Recovery Amount}}{\text{Default Amount}}$

Is more useful as it describes what percentage of the debt is recovered. The average recovery rate in this data set is 0.42 (not including debts still being paid). Some debts have negative recovery rate, which is because default amounts generate interests in the following months after default, but the debtors did not pay anything, so the outstanding balance keeps increasing. These are redefined to be 0. Some debts have recovery rate greater than 1, which occur when the debtors paid back the entire amount at default and also the interest and collection fees which was subsequently charged on it. For these cases, the recovery rates are redefined to be 1.

The distribution of recovery rate is a bathtub shape, see figure (2). 30.3% debts have 0 recovery rate, and 23.9% debts have 100% recovery rate, others are relatively evenly distributed between 0 and 1. (The distribution excludes the debts still being paid.)

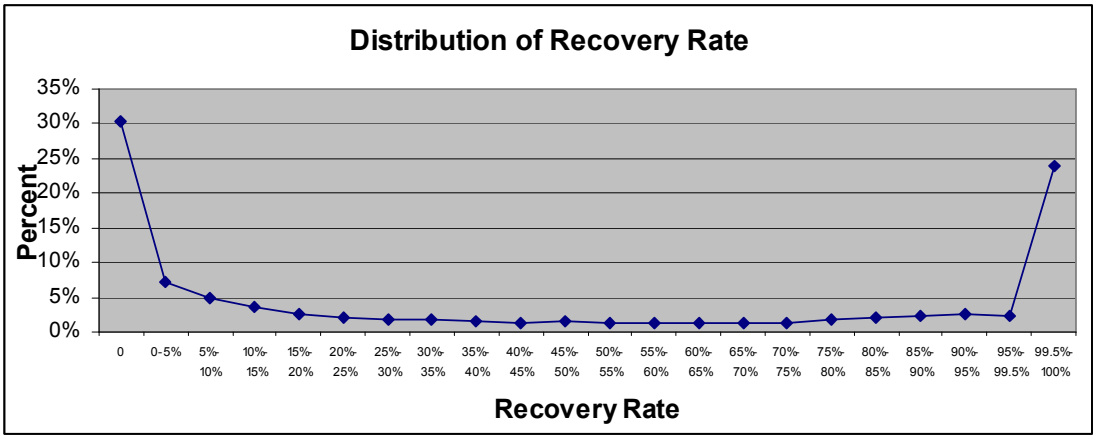

Figure 2: Distribution of recovery rate in the data set

The whole data is randomly split into 2 parts; the training sample contains 70% of observations for building models, and the test sample contains 30% of observations for testing and comparing models.

In the following sections, the modelling details will be presented. Results from linear regression and survival analysis models will be compared; and also comparison between results from single distribution models and mixture distribution models will be made.

4.2 Single distribution models

Linear regression

Two multiple linear regression models are built, one is for recovery rate as target variable and one is for recovery amount as target variable. In the former case, the

predicted recovery rate could be multiplied by default amount, thus the recovery amount could be predicted indirectly; in the latter case, the predicted recovery rate was obtained by dividing predicted recovery amount by default amount.

The stepwise selection method was set for regression models. Coarse classification was used on categorical variables with attributed with similar average target variable values put in the same class. The two continuous variables 'default amount' and 'ratio of default amount to total loan' were transformed into ordinal variables as well, and also their functions (square root, logarithm, and reciprocal) and original form were included in the model building in order to better fit the Recovery Rate.

The R-squares for these models are small, which is consistent with previous authors, but they are statistically significant. The Spearman rank correlation reflects how accurate was the ranking of the predicted values. From the results table (1), we can see modelling recovery rate directly is better than indirect modelling from recovery amount model, and better recovery amount results are also obtained by predicting recovery rate first.

	R-square	Spearman	MAE	MSE
Recovery Rate from recovery rate model	0.1066	0.3183	0.3663	0.1650
Recovery Rate from recovery amount model	0.0354	0.2384	0.4046	0.2352
Recovery Amount from recovery amount model	0.1968	0.2882	1239.2	2774405.4
Recovery Amount from recovery rate model	0.2369	0.3307	1179.6	2637470.7

Table 1: Linear regression models (results were from training sample)

Table 1 lists the results of linear regression models from training sample. In the recovery rate modelling, the most significant variable is 'the ratio of default amount to total loan', which has a negative relation with recovery rate. This gives some indication of how much of the loan was still owed before default occurs, and if a substantial portion of the loan was repaid before default then the Recovery Rate is also likely to be high. The second most significant variable is 'second applicant status'. The model results show loans with second applicant have higher recovery rate than loans without second applicant, maybe because there is a second potential income stream to help pay the recovery. Other significant variables include: employment status, residential status, and default amount. The model details can be found in Table 2. In the recovery amount model, the variables which entered the model are very similar to recovery rate model. Because predicted recovery amount from recovery amount model is worse than that from the recovery rate model, the coefficient details of recovery amount model are not given in this paper.

Survival analysis

There are two reasons why survival analysis could be used. First, some loans in the data set are still being paid; these observations can not be included in the linear regression model. Survival analysis models can treat them as censored, and include them in model building. Second, recovery rate is not normally distributed; in certain sense, linear regression violates its assumption. Survival analysis models can handle

this problem; different distributions can be set in accelerated models and Cox model's approach allows any empirical distribution.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	43	321.28829	7.47182	45.17	<.0001
Error	16283	2693.34350	0.16541		
Corrected Total	16326	3014.63178			

Root MSE	0.40670	R-Square	0.1066
Dependent Mean	0.42323	Adj R-Sq	0.1042
Coeff Var	96.09587		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	0.68238	0.02856	23.89	<.0001	0
emp1	1	0.09759	0.01257	7.76	<.0001	2.55646
emp2	1	0.14370	0.01487	9.66	<.0001	2.50773
mort	1	0.04706	0.00870	5.41	<.0001	1.47386
visa	1	-0.03558	0.00993	-3.58	0.0003	1.07530
ind2	1	-0.05303	0.00877	-6.05	<.0001	1.86393
dep2	1	0.02678	0.01019	2.63	0.0086	1.04863
pl	1	0.02392	0.00769	3.11	0.0019	1.31492
resi1	1	-0.03681	0.01057	-3.48	0.0005	1.27696
resi3	1	-0.04130	0.01694	-2.44	0.0148	1.09436
resi4	1	-0.11322	0.01270	-8.92	<.0001	1.43363
sav	1	0.01422	0.00675	2.11	0.0351	1.10852
term1	1	-0.06325	0.01863	-3.40	0.0007	1.34068
term2	1	-0.02683	0.01012	-2.65	0.0080	1.43729
term4	1	0.04176	0.01103	3.79	0.0002	1.05892
app21	1	-0.10711	0.01373	-7.80	<.0001	1.94432
app22	1	-0.05145	0.01700	-3.03	0.0025	1.18274
app23	1	-0.12671	0.00872	-14.53	<.0001	1.72156
purp1	1	-0.06913	0.01575	-4.39	<.0001	1.16256
purp2	1	-0.03990	0.00875	-4.56	<.0001	1.42328
purp3	1	-0.05139	0.01232	-4.17	<.0001	1.15933
purp4	1	-0.04359	0.00963	-4.53	<.0001	1.30166
ad2	1	0.03269	0.01098	2.98	0.0029	1.07328
ad3	1	0.03659	0.01009	3.63	0.0003	1.12161
ad4	1	0.05110	0.01334	3.83	0.0001	1.10450
ad5	1	0.06550	0.01469	4.46	<.0001	1.08694
ad6	1	0.07402	0.01495	4.95	<.0001	1.12889
ad7	1	0.08950	0.01404	6.37	<.0001	1.13754
ha1	1	-0.03007	0.01466	-2.05	0.0403	1.21402
ha5	1	0.03238	0.01034	3.13	0.0017	1.04398
oc1	1	0.02873	0.01298	2.21	0.0268	1.03445
oc2	1	0.03932	0.01298	3.03	0.0025	1.04891
oc3	1	0.04387	0.01509	2.91	0.0037	1.04048
oc4	1	0.04660	0.01520	3.07	0.0022	1.04525
oc5	1	0.09007	0.01574	5.72	<.0001	1.06338
exp1	1	0.03618	0.01558	2.32	0.0202	2.01901
income1	1	0.06619	0.01275	5.19	<.0001	3.68252
income2	1	0.05957	0.01282	4.65	<.0001	1.68589
afford3	1	0.05652	0.01572	3.60	0.0003	1.59263
def_year90	1	0.03107	0.01012	3.07	0.0021	1.21026
def_year96	1	0.02887	0.01083	2.67	0.0077	1.19851
srt_default	1	-0.00276	0.00028412	-9.72	<.0001	3.15912
rec_default	1	-58.39810	8.93313	-6.54	<.0001	2.64808
doo	1	-0.01246	0.00068373	-18.22	<.0001	1.54152

'emp': employment status; 'mort': with mortgage; 'visa': with visa card; 'ind': insurance indicator; 'dep': number of dependants; 'pl': with personal loan account; 'resi': residential status; 'sav': with saving account; 'term': loan term; 'app2': second applicant status; 'purp': loan purpose; 'ad': time at address; 'ha': time with the bank; 'oc': time in occupation; 'exp': monthly expenditure; 'income': monthly income; 'afford': the ratio of expenditure to income; 'def_year': default year; 'srt_default': square root of default amount; 'rec_default': reciprocal of default amount; 'doo': ordinal variable of the ratio of default amount to total loan.

Table 2: Coefficients of variables in single distribution linear regression model for recovery rate:

Both accelerated failure time models and proportional hazards models (Cox regression model) are built for modelling both recovery rate and recovery amount. Here, the event of interest is debt write off, so the write-off debts are treated as uncensored; debts which were paid off or were still being paid are treated as censored. All the independent variables which are used in the linear regression model building are used here as well, and they are regrouped into dummy variables. Continuous variables were firstly cut into 10 to 15 bins to become 10 to 15 dummy variables, and put them into survival analysis model without any other characteristics. Observing coefficients of them from model output and bins with similar coefficients were binned together with their neighbours. The same method was used for nominal variables. Two continuous variables 'default amount' and 'ratio of default amount to total loan' were included in the models as their original forms as well.

Because accelerated failure time models can not handle 0's existing in target variable, observations with recovery rate 0 should be removed off from the training sample before building the accelerated failure time models. This leads to a new task: a classification model is needed to classify recovery 0's and non-0's (recovery rate greater than 0). Therefore, a logistic regression model is built based on training sample before building accelerated failure time models. In the logistic regression model, the variables 'month until default' and 'loan term' are very significant, which are not so important in the linear regression models before, other variables selected into the model are similar to previous regression models. The Gini coefficient is 0.32 and 57.8% 0's were predicted as non-0's and 21.5% non-0's were predicted as 0's by logistic regression model. Cox regression model can allow 0's to exist in target variable; so two Cox models are built, one including 0 recoveries and another excluding.

For the accelerated failure life models, the type of distribution of survival time needs to be chosen. After some simple distribution tests, Weibull, Log-logistic and Gamma distributions are chosen for recovery rate model; and Weibull and Log-logistic distributions are chosen for recovery amount model. Cox model is called semi-parametric, and there is no need to concern which family of distribution to use.

Recovery Rate	Optimal quantile	Spearman	MAE	MSE
Accelerated (Weibull)	34%	0.24731	0.3552	0.1996
Accelerated (log-logistic)	34%	0.25454	0.3532	0.2015
Accelerated (gamma)	36%	0.16303	0.3597	0.1968
Cox-with 0 recoveries	46%	0.24773	0.3631	0.2092
Cox-without 0 recoveries	30%	0.24584	0.3604	0.2100

Table 3: Survival analysis models results for recovery rate

Unlike linear regression, survival analysis models generate a whole distribution of the predicted values for each debt, rather than a precise value. Thus, to give a precise value, the quantile or mean of the distribution can be considered. In all the survival models, the mean and median values are not good predictors, because they are too big

and generate large MAE and MSE compared with predictions from some other quantiles. The optimal predicting quantile points are chosen based on minimum MAE and/or MSE. The lowest MAE and MSE are found with quantile levels lower than median, and the results from the training sample models are listed in Table 3 and Table 4. The model details of Cox-with 0 recoveries can be found in Table 5.

Recovery Amount	Optimal quantile	Spearman	MAE	MSE
Accelerated (Weibull)	34%	0.30768	1129.7	3096952
Accelerated (log-logistic)	34%	0.31582	1117.0	3113782
Cox-with 0 recoveries	46%	0.29001	1174.5	3145133
Cox-without 0 recoveries	30%	0.30747	1140.25	3112821

Table 4: Survival analysis models results for recovery amount

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
mort	1	-0.14199	0.02404	34.8840	<.0001	0.868
visa	1	0.10609	0.02743	14.9529	0.0001	1.112
pl	1	-0.08675	0.02149	16.2972	<.0001	0.917
remp1	1	-0.07858	0.04003	3.8525	0.0497	0.924
remp2	1	0.06436	0.03279	3.8529	0.0497	1.066
remp3	1	0.32792	0.04487	53.4124	<.0001	1.388
rind2	1	0.09904	0.02973	11.0976	0.0009	1.104
rind3	1	0.11452	0.03190	12.8890	0.0003	1.121
rmari2	1	0.08955	0.03090	8.3959	0.0038	1.094
rdep	1	-0.06428	0.02132	9.0893	0.0026	0.938
rresi1	1	0.09188	0.02896	10.0630	0.0015	1.096
rresi3	1	0.26498	0.02947	80.8485	<.0001	1.303
rapp21	1	-0.22500	0.02521	79.6685	<.0001	0.799
rapp22	1	-0.14470	0.04558	10.0800	0.0015	0.865
rpurp1	1	0.14572	0.02207	43.5912	<.0001	1.157
rpurp2	1	0.12971	0.02588	25.1156	<.0001	1.138
rage	1	-0.05094	0.02382	4.5729	0.0325	0.950
rad	1	-0.16348	0.02336	48.9907	<.0001	0.849
roc	1	-0.14658	0.02381	37.9037	<.0001	0.864
rha1	1	-0.05968	0.02342	6.4968	0.0108	0.942
rha2	1	-0.11485	0.02989	14.7607	0.0001	0.892
rha3	1	-0.21461	0.03078	48.6113	<.0001	0.807
rafford	1	0.16975	0.03110	29.7880	<.0001	1.185
rdoo1	1	0.09040	0.02675	11.4192	0.0007	1.095
rdoo2	1	0.18290	0.02827	41.8741	<.0001	1.201
rdoo3	1	0.32359	0.03881	69.5068	<.0001	1.382
rdoo4	1	0.34011	0.05030	45.7239	<.0001	1.405
rdoo5	1	0.43940	0.05210	71.1407	<.0001	1.552
rdef1	1	0.11177	0.04361	6.5678	0.0104	1.118
rdef3	1	-0.06819	0.02672	6.5141	0.0107	0.934
rdef4	1	0.05920	0.02709	4.7748	0.0289	1.061
rdef5	1	0.18287	0.04071	20.1786	<.0001	1.201
rdef6	1	0.20961	0.04416	22.5273	<.0001	1.233
rmon1	1	0.12048	0.03893	9.5769	0.0020	1.128
rmon2	1	0.06726	0.02702	6.1954	0.0128	1.070
ldef91	1	0.10103	0.02738	13.6107	0.0002	1.106
ldef92	1	0.08227	0.03845	4.5776	0.0324	1.086
ldef93	1	0.11611	0.04528	6.5759	0.0103	1.123
ldef95	1	-0.10487	0.05003	4.3928	0.0361	0.900
ldef96	1	-0.20252	0.04420	20.9925	<.0001	0.817
ldef97	1	-0.18982	0.04567	17.2769	<.0001	0.827
ldef98	1	-0.21648	0.04586	22.2866	<.0001	0.805
ldef99	1	-0.16531	0.06395	6.6822	0.0097	0.848

'mort': with mortgage; 'visa': with visa card; 'pl': with personal loan account; 'remp': employment status; 'rind': insurance indicator; 'rdep': number of dependants; 'rmari': marital status; 'rresi': residential status; 'rapp2': second applicant status; 'rpurp': loan purpose; 'rage': age when applying; 'rad': time at address; 'rha': time with the bank; 'roc': time in occupation; 'afford': the ratio of expenditure to income; 'doo': ordinal variable of the ratio of default amount to total loan; 'rdef': default amount; 'rmon': month until default; 'ldef': default year;

Table 5: Coefficients of variables in single distribution Cox regression model (including 0 recovery) for recovery rate:

Using a quantile value has some advantages in this case and quantile regression has been applied in credit scoring research. Whittaker et al (2005) use quantile regression to analyse collection actions, and Somers and Whittaker (2007) use quantile regression for modelling distributions of profit and loss. Benoit and Van den Poel (2009) apply quantile regression to analyse customer life value. Using quantile values to make prediction can avoid outlier influence. In particular when using survival analysis, the mean value of a distribution is affected by the amount of censored observations in the data set, so use a quantile value is a good idea to make predictions.

If the Spearman rank correlation test is the criterion to judge the model, we can see, from the above results tables (table2 and table3), the accelerated failure time model with log-logistic distribution is the best one among several survival analysis models. We can also see the optimal quantile point is almost the same regardless of the distribution in accelerated failure time models. The number of censored observations in the training sample does influence the optimal quantile point. If some of the censored observations are deleted from the training sample, the optimal quantile points move towards the median.

Model comparison

Model comparison is made based on test sample. For some debts still being paid, the final recovery amount and recovery rate are not known, and they can't be measured properly, thus these observations are removed from the test sample. All the predicted results from single distribution models are listed in Tables 6 and 7:

Recovery Rate	R-square	Spearman	MAE	MSE
(1) Linear Regression	0.0904	0.29593	0.3682	0.1675
(2) A – Weibull	0.0598	0.25306	0.3586	0.2042
(3) A – log-logistic	0.0638	0.25990	0.3560	0.2060
(4) A – gamma	0.0527	0.23496	0.3635	0.2015
(5) Cox – including 0's	0.0673	0.27261	0.3546	0.2006
(6) Cox – excluding 0's	0.0609	0.25506	0.3564	0.2072
(7) Linear Regression*	0.0292	0.22837	0.4077	0.2432
(8) A – weibull*	0.0544	0.24410	0.3606	0.2070
(9) A – log-logistic*	0.0591	0.25315	0.3575	0.2077
(10) Cox – including 0's*	0.0425	0.22646	0.3693	0.2216
(11) Cox – excluding 0's*	0.0504	0.23269	0.3624	0.2108

*: results from recovery amount models

Table 6: Comparison of recovery rate from single distribution models test sample

From the recovery rate Table 6, if R-square and Spearman ranking test are the criterion to judge a model, we can see (1) Linear Regression is the best one, and (5) Cox-including 0's is the second best model. In the training sample, accelerated failure time model with log-logistic distribution outperforms the Cox models, but for the test sample, Cox model including 0's is more robust than accelerated failure models. In terms of MSE, linear regression always achieves the lowest MSE as one would expect to see it is essentially minimising that criterion. All the survival models have similar results. For MAE, the results are more consistent, except the linear regression models are poor. It is also can be noticed that to model recovery rate directly is better than to model recovery rate from recovery amount models. Almost all the R-square and

Spearman test from recovery amount models are lower than these from recovery rate models.

Recovery Amount	R-square	Spearman	MAE	MSE
(1) Linear Regression	0.1807	0.28930	1212.1	2634270
(2) A – weibull	0.1341	0.30594	1123.5	3026908
(3) A – log-logistic	0.1318	0.31178	1111.7	3047317
(4) Cox – including 0's	0.1572	0.31788	1138.9	2887499
(5) Cox – excluding 0's	0.1400	0.30437	1125.3	3017661
(6) Linear Regression*	0.2068	0.32522	1162.4	2549591
(7) A – weibull*	0.1424	0.31149	1116.1	2982477
(8) A – log-logistic*	0.1396	0.31697	1105.9	3014320
(9) A – gamma*	0.1413	0.30139	1141.5	2972807
(10) Cox – including 0's*	0.1628	0.34619	1101.9	2906821
(11) Cox – excluding 0's*	0.1377	0.31246	1107.4	3028183

*: results from recovery rate models

Table 7: Comparison of recovery amount from single distribution models test sample

From recovery amount Table 7, we see that modelling recovery amount directly is not as good as estimating recovery rate first, because (6) Linear Regression* model achieves the highest R-square and (10) Cox-including 0's* model achieves the highest Spearman ranking coefficient. Both of them are recovery rate models and the predicted recovery amount is calculated by multiplying predicted recovery rate by the default amount. Regression models and Cox-including 0's models outweigh the accelerated failure time models.

In the training sample, we can see Cox models and Accelerated failure time models have very similar results (in terms of spearmen rank test). In the test sample, Cox-including 0's model beats the other survival models. The reason is that the logistic regression model which is used before the other models to classify 0 recoveries and non-0 recoveries generates more errors in the test sample, but Cox-including 0's model is not affected by this.

5 Mixture distribution models

Mixture distribution models have the potential to improve prediction accuracy and they have been investigated by other researchers for modelling RR. Thomas et al (2007) suggested to separate LGD=0 and LGD>=0 for unsecured personal loans, and then modelling LGD by implementing different collection actions. Bellotti and Crook (2009) suggested to separate RR=0, 0<RR<1, and RR=1 for credit cards, and then for the group 0<RR<1, use OLS or LAV regression to model RR and achieved R-square 0.077. One reason of separately modelling RR is people's different views about repayment. Some debtors want to pay back, but they have financial troubles and can't pay back; but some debtors deliberately do not want to pay.

Method 1

The recovery rate is treated as a continuous variable and also the target variable, and a classification tree model is built to split the whole population into a few subgroups, in order to maximise the difference of average recovery rate among them.

As is seen from the tree in Figure 3, the whole population is split into 4 segments in the end nodes. Generally, large amount loans have lower recovery rate than small amount loans; if the debtors have mortgage in this bank, then their loans have higher recovery rate than those without mortgage in the bank; people of house owners or living with parents have higher recovery rate than people of tenets or other residential status.

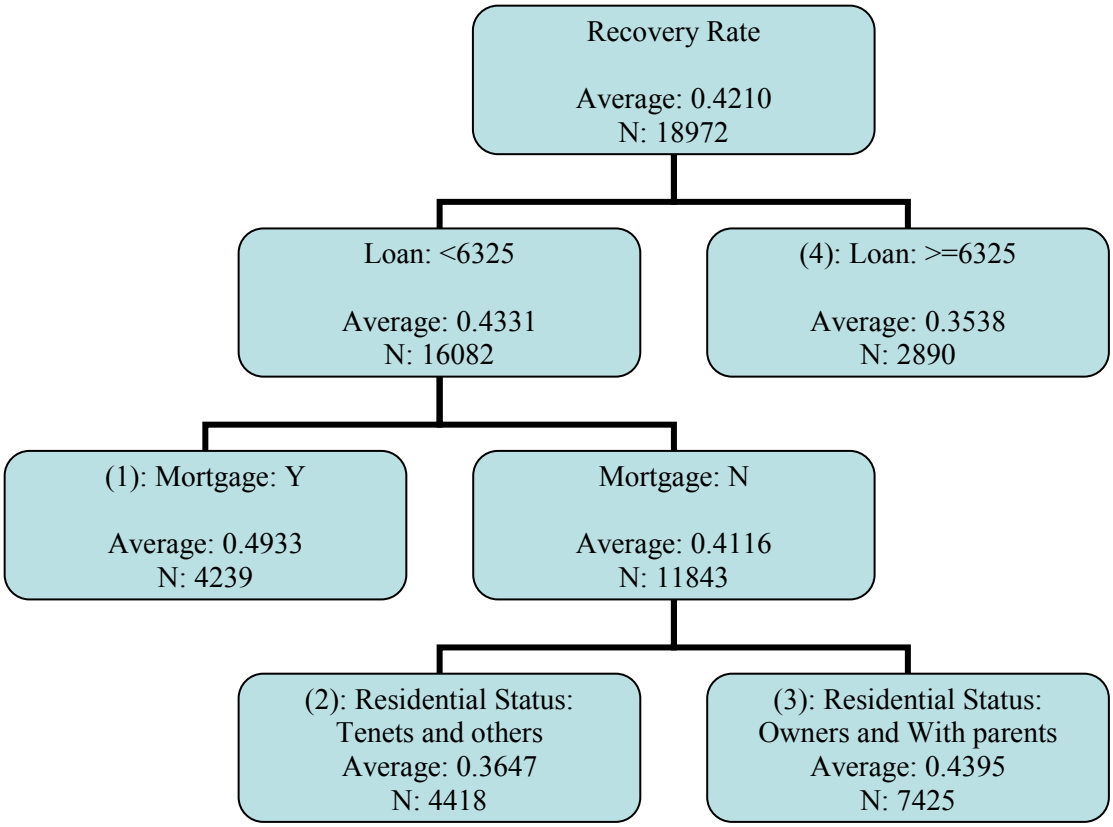

Figure 3: Classification tree for recovery rate as continuous variable

Linear regression model and survival models are built for each of the segments. The previous research shows that better predicted recovery amount results are obtained from predicting recovery rate first and then multiplying by default amount, so only recovery rate models are built here. The models are built based on training samples and tested on holdout samples.

Recovery Rate	R-square	Spearmen	MAE	MSE
Regression	0.0840	0.28544	0.3693	0.1688
Accelerated	0.0660	0.26625	0.3549	0.2055
Cox-including 0's	0.0752	0.28581	0.3518	0.1967
Cox-excluding 0's	0.0636	0.26236	0.3549	0.2067

Table 8: Recovery rate from mixture distribution models of method 1

In all four segments, linear regression is always the best modelling technique, as it has the highest R-square and Spearmen coefficient; so after piecing together the 4 segments, linear regression model still has the highest R-square. In the training samples of accelerated failure time models, the first 3 segments achieve better results (higher R-square and spearmen rank coefficient) in log-logistic distribution models, and the last segment has a better result using the Weibull distribution model. So the

test results for the accelerated failure time models are made up of three log-logistic distribution models and one Weibull distribution model. In the Cox-regression modelling, the Cox model including 0's (without logistic regression to predict 0 or non-0 recoveries) performs better than Cox model excluding 0's (with logistic regression first) in all four subgroups. This means it is not better to predict 0 recoveries by logistic regression first.

Recovery Amount	R-square	Spearmen	MAE	MSE
Regression	0.1942	0.31824	1166.7	2593870
Accelerated	0.1346	0.31820	1102.3	3030185
Cox-including 0's	0.1574	0.35314	1100.5	2976283
Cox-excluding 0's	0.1357	0.31564	1105.8	3068188

Table 9: Recovery amount from mixture distribution models of method 1

In terms of R-square, among mixture distribution models, the linear regression models are the best; but in terms of spearmen ranking test, the Cox model-including 0's outperforms the linear regression model, especially for predicting recovery amount.

Compared with the analysis from single distribution models, the results from mixture distribution models are disappointing and are almost all worse than results from single distribution models. In terms of R-square, the best model in mixture distribution models is linear regression, but its R-square is still lower than that from single distribution linear regression model. In terms of Spearmen ranking coefficient, the best model in mixture distribution models is Cox model-including 0's. The Spearmen ranking coefficient for recovery rate is a little bit lower than 0.29593 which is the best one in the single distribution models; the Spearmen ranking coefficient for recovery amount is higher than 0.34619 which is the highest in the single distribution models. Thus, this mixture distribution models only improve the spearmen rank coefficient in the case of recovery amount predictions.

Method 2

Another way to separate the whole population is to split the target variable into three groups: the first group $RR<0.05$ (almost no recoveries), the second group $0.05<RR<0.95$ (partial recoveries), and the third group $RR>0.95$ (full recoveries). These splits correspond to essentially no, partial or full recovery and there is a belief that a particular defaulter is most likely to be in one of these group because of his circumstance.

Recovery rate can be treated as an ordinal variable, with three classes - recovery rate less than 0.05 is set to 0, recovery rate between 0.05 and 0.95 is set 1, and recovery rate greater than 0.95 is set 2. A classification tree with the three classes as the target variable was tried, but the results were disappointing because each end node had similar distribution over the three classes. As an alternative a classification tree was first built to separate 0's and non-0's, so the whole data is split into two groups; secondly. Then a second classification tree was built for the non-0's group, in order to separate them into 1's and 2's. So again the population was split into 3 subgroups and this gave slightly better results. The population in the first segment (most zero repayments) have the following attributes: no mortgage and loan term less than or equal to 12 months, OR no mortgage, time at address less than 78 months and have a current account. The population in the third segment (highest full repayment rate)

have attributes: loan less than £4320 and insurance accepted. The rest of the population are allocated to the second segment.

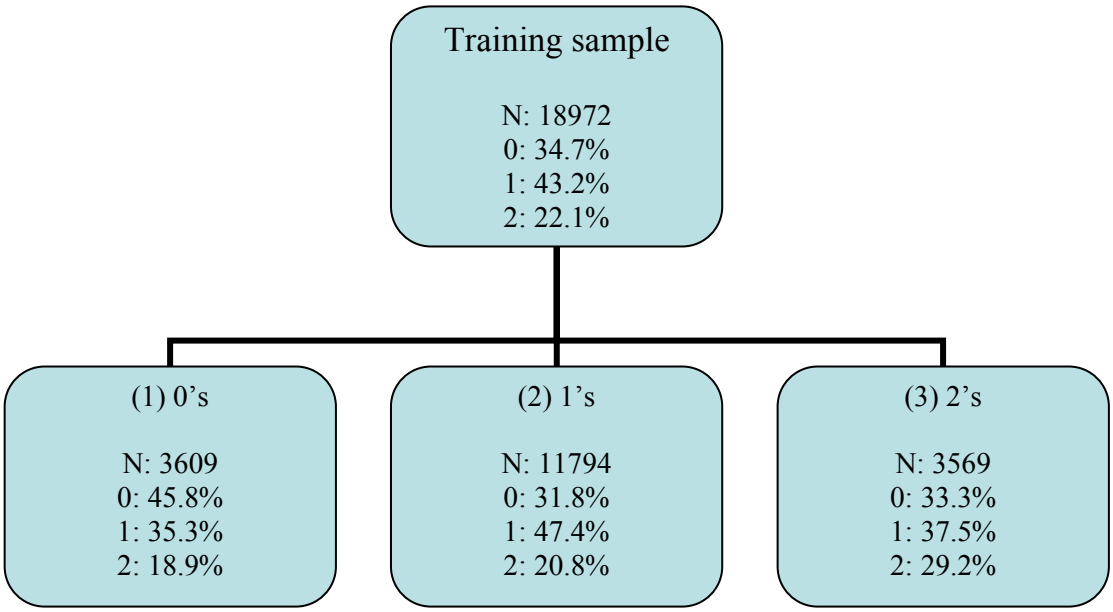

Figure 4: Classification tree for recovery rate as ordinal variable

This classification is very coarse. Group (1) aims at debts with recovery rate less than 0.05, but only 45.8% debts actually belong to this group; group (2) is for the debts with recovery rate between 0.05 and 0.95, but only 47.4% debts are in this range; group (3) is for the debts with recovery rate greater than 0.95, but, only 29.2% debts in this group have recovery rate greater than 0.95.

In the previous analysis, the linear regression model and Cox-including 0's model are the two best models, so here only the linear regression model and the Cox-including 0's regression model are built for each of the three segments. The results from the combined test sample are compared with the results from previous research, see Tables 10 and 11.

Recovery Rate	R-square	Spearmen	MAE	MSE
Regression	0.0734	0.26453	0.3695	0.1688
Cox including 0's	0.0570	0.25869	0.3588	0.2051

Table 10: Recovery rate from mixture distribution models of method 2

Recovery Amount	R-square	Spearmen	MAE	MSE
Regression	0.2054	0.31356	1169.4	2564149
Cox including 0's	0.1669	0.33888	1125.7	2930725

Table 11): Recovery amount from mixture distribution models of method 2

From Tables 10 and 11, we can see that, for recovery rate, the linear regression model is still better than the Cox regression model in terms of R-square and Spearmen coefficient; for recovery amount, the R-square of linear regression model is higher than that of the Cox regression model, but the Spearmen coefficient of linear regression is lower than that of the Cox model. Compared with the analysis results

from single distribution models, this mixture model seems does not improve the R-square or the Spearmen ranking coefficient.

5 Conclusions

Estimating Recovery Rate and Recovery Amount has become much more important because of both the new Basel Accord regulation and because of the increase in the number of defaulters due to the recession.

This paper makes a comparison between single distribution and mixture distribution models of predicting recovery rate for unsecured consumer loans. Linear regression and survival analysis are the two main techniques used in this research.. Linear regression can model recovery rate and recovery amount directly; accelerated failure time models do not allow 0's to exist in the target variable, so a logistic regression model is built first to classify which loans have zero and which have non zero recovery rates. Cox's proportional hazard regression models can deal with 0's in the target variable, and so that approach was tried both with logistic regression used first to split off the zero recoveries and without using logistic regression first.

In the comparison of the single distribution models, the research result shows that linear regression is better than survival analysis models in most situations. For recovery rate modelling, linear regression achieves higher R-square and Spearman rank coefficient than survival analysis models. The Cox model without logistic regression first is the best model among all the survival analysis models. For recovery amount modelling, it is better to predict recovery amount from a recovery rate model than to model it directly.

This is surprising given the flexibility of distribution that the Cox approach allows. Of course one would expect MSE to be minimised using linear regression on the training sample because that is what linear regression tries to do. However, the superiority of linear regression holds for the other measures both on the training and the test set. One reason may be the need to split off the zero recovery rate cases in the second analysis approach. This is obviously difficult to do and the errors from this first stage results in a poorer model in the second stage. This could also be the reason that the mixture models do not give a real improvement. Of course, one could choose the mixtures using other characteristics as well as the recovery rate which is used here.

Another reason for the survival analysis approach not doing so well in this data set is the there is only a relatively small amount of data where payment is still going on (14%). This is because the data is relatively old and has been held for a long period. A data set which more up to date would have a large proportion of loans still repaying. Lastly, there is the question of whether loans with RR=1 are really censored or not. Assuming they are not censored would lead to model lower estimate of RR, which might be more appropriate for the consolidate philosophy of the Basel Accord.

References:

Altman E., Eberhart A., (1994), Do Seniority Provisions protect bondholders' investments, J. Portfolio Management, Summer, pp67-75

Altman E., Haldeman R., Narayanan P., (1977), ZETA Analysis: A new model to identify bankruptcy risk of corporations, Journal of banking and Finance 1, pp29-54

Altman E. I., Resti A., Sironi A.: Analyzing and Explaining Default Recovery Rates A Report Submitted to The International Swaps & Derivatives Association, December 2001

Altman E. I., Resti A., Sironi A. (2005): Loss Given Default; a review of the literature in Recovery Risk, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, pp 41-59

Basel Committee on Banking Supervision (BCBS), (2004, updated 2005), International Convergence of Capital Measurement and Capital standards: a revised framework, Bank of International Settlement, Basel.

Bellotti T., Crook J., (2009) Calculating LGD for Credit Cards, presentation in Conference on Risk Management in the Personal Financial Services Sector, London, 22-23 January 2009
http://www3.imperial.ac.uk/mathsinstitute/programmes/research/bankfin/qfrmc/events/past/jan09conference

Benoit D.F., Van den Poel D. (2009) Benefits of quantile regression for the analysis of customer lifetime value in a contractual setting: An application in financial services, *Expert Systems with Applications* 36 (2009) 10475-10484

Chew W.H., Kerr S.S., (2005), Recovery Ratings: Fundamental Approach to Estimation Recovery Risk, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p87-97

Cox D.R., (1972), Regression Models and Life Tables (with discussion), *Journal of the Royal Statistical Society*, B34, 187-220

De Servigny A., Oliver R., (2004), Measuring and managing Credit Risk, McGraw Hill, Boston

Dermine J., Neto de Carvalho C., (2006), Bank loan losses given default: A case study, Journal of banking and Finance 30, 1219-1243.

Dias Jose G. (2004), *Finite Mixture Models*, Rijksuniversiteit Groningen

Engelmann B., Rauhmeier R., (2006), The Basel II Risk Parameters, Springer, Heidelberg

Gupton G. (2005), Estimation Recovery Risk by means of a Quantitative Model: LossCalc, *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p61-86

Lucas A.: Basel II Problem Solving; QFRMC Workshop and conference on Basel II & Credit Risk Modelling in Consumer Lending, Southampton 2006;

McLachlan G. J., Basford K. E. (1998), *Mixture Models: Inference and Applications to Clustering*, New York: Marcel Dekker.

Matuszyk A., Mues C, Thomas L.C., Modelling LGD for unsecured personal loans: Decision tree approach, Working paper CORMSIS, School of Management, University of Southampton; to appear in Journal of Operational Research Society on-line.

Qi M., Yang X., (2009), Loss given default of high loan-to-value residential mortgages, *Journal of Banking & Finance* 33 (2009) 788-799

Querci F., (2005) Loss Given Default on a medium-sized Italian bank's loans: an empirical exercise, The European Financial Management Association, Genoa, Genoa University. http://www.efmaefm.org/efma2005/papers/206-querci_paper.pdf

Schuermann T., (2005), What Do We Know About Loss Given Default? *Recovery Risk*, ed by Altman E.I., Resti A, Sironi A. Risk books, London, p3-24

Somers M., Whittaker J. (2007) Quantile regression for modelling distributions of profit and loss, European Journal of Operational Research 183 (2007) 1477-1487

Wedel M., Kamakura W. A., (2000), *Market Segmentation. Conceptual and Methodological Foundations* (2nd ed.), International Series in Quantitative Marketing, Boston: Kluwer Academic Publishers.

Whittaker J., Whitehead C., Somers M., (2005). The neglog transformation and quantile regression for the analysis of a large credit scoring database. *Applied Statistics-Journal of the Royal Statistical Society* Series C 54, 863–878.